Language models can learn complex molecular distributions

Year: 2021 | Citations: 169 | Authors: Daniel Flam-Shepherd, Kevin Zhu, A. Aspuru■Guzik

Abstract

Deep generative models of molecules have grown immensely in popularity, trained on relevant datasets, these models are used to search through chemical space. The downstream utility of generative models for the inverse design of novel functional compounds, depends on their ability to learn a training distribution of molecules. The most simple example is a language model that takes the form of a recurrent neural network and generates molecules using a string representation. Since their initial use, subsequent work has shown that language models are very capable, in particular, recent research has demonstrated their utility in the low data regime. In this work, we investigate the capacity of simple language models to learn more complex distributions of molecules. For this purpose, we introduce several challenging generative modeling tasks by compiling larger, more complex distributions of molecules and we evaluate the ability of language models on each task. The results demonstrate that language models are powerful generative models, capable of adeptly learning complex molecular distributions. Language models can accurately generate: distributions of the highest scoring penalized LogP molecules in ZINC15, multi-modal molecular distributions as well as the largest molecules in PubChem. The results highlight the limitations of some of the most popular and recent graph generative models– many of which cannot scale to these molecular distributions. Generative models for the novo molecular design attract enormous interest for exploring the chemical space. Here the authors investigate the application of chemical language models to challenging modeling tasks demonstrating their capability of learning complex molecular distributions.